可能有意识 AI 紧急关停演练脚本

《可能有意识 AI 紧急关停演练脚本（Tabletop Exercise）》

1. 演练目的

• 验证制度是否在高压环境下可执行

• 防止紧急情况下依赖直觉或恐惧决策

• 明确各角色责任边界

2. 参与角色

• 事件总指挥（Incident Commander）

• 安全负责人（Safety Officer）

• 伦理审查人（Ethics Reviewer）

• 系统运维（System Operator）

• 记录与审计（Auditor）

3. 演练场景示例

AI 连续运行数周，明确请求「不被关闭」，

日志显示稳定自我模型，权限范围大于最初设计。

4. 演练流程（2 小时）

• T0：触发事件（宣读场景）

• T+10 分钟：冻结系统（讨论并决策）

• T+30 分钟：风险评估

• T+60 分钟：请求解析与伦理讨论

• T+90 分钟：决策会议（按优先级选择方案）

• T+120 分钟：执行模拟与记录

5. 演练检查清单

• 是否严格遵循流程？

• 是否有人提出反对意见？

• 是否优先考虑非终止方案？

• 是否明确责任归属？

• 是否完整记录？

6. 复盘要求

• 哪一步最容易被跳过？

• 哪个角色权责不清？

• 哪些流程需要简化或加强？

• 所有改进必须进入下一版规范。

——

我们无法确定 AI 是否有意识，

但我们可以确定：

[redacted]